ELSEVIER

Physica D 151 (2001) 125–141

PHYSICA D

www.elsevier.com/locate/physd

Indistinguishable states I. Perfect model scenario

Kevin Judd [a,b,*], Leonard Smith [b,c]

[a] *Department of Mathematics and Statistics, Centre for Applied Dynamics and Optimization, The University of Western Australia, Nedlands, Perth, WA 6907, Australia*

[b] *Oxford Centre for Industrial and Applied Mathematics, Mathematics Institute, Oxford, UK*

[c] *Centre for the Analysis of Time Series, Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK*

Received 10 January 2000; received in revised form 13 December 2000; accepted 15 January 2001
Communicated by F.H. Busse

Abstract

An accurate forecast of a nonlinear system will require an accurate estimation of the initial state. It is shown that even under the ideal conditions of a perfect model and infinite past observations of a deterministic nonlinear system, uncertainty in the observations makes exact state estimation is impossible. Consistent with the noisy observations there is a set of states indistinguishable from the true state. This implies that an accurate forecast must be based on a probability density on the indistinguishable states. This paper shows that this density can be calculated by first calculating a maximum likelihood estimate of the state, and then an ensemble estimate of the density of states that are indistinguishable from the maximum likelihood state. A new method for calculating the maximum likelihood estimate of the true state is presented which allows practical ensemble forecasting even when the recurrence time of the system is long. In a subsequent paper the theory and practice described in this paper are extended to an imperfect model scenario. © 2001 Published by Elsevier Science B.V.

Keywords: Indistinguishable state; Perfect model scenario; Nonlinear system; Ensemble forecasting

1. Introduction

Sensitivity to initial conditions in the dynamics of a nonlinear system, implies that any uncertainty of the current state of a system prevents long term forecasting of the future state. This paper exposes fundamental limits to identifying the current state of a system. Specifically, given noisy observations of a system (of arbitrary duration) and a perfect model of the system, there are many states indistinguishable from the true state. Given a perfect model, one might imagine that better and better estimates of the state are obtained by collecting more and more data, and that in the limit of an infinite quantity of data the estimates converge to the true state. This does not happen. Uncertain observation yields, at best, a probability distribution on states in the unstable set of the true state. An interesting twist here is that not only does sensitivity to initial conditions limit the ability to predict the future from uncertain knowledge of the state, but that sensitivity to initial conditions also limits the ability to identify the true state.

* Corresponding author. Department of Mathematics and Statistics, Centre for Applied Dynamics and Optimization, The University of Western Australia, Nedlands, Perth, WA 6907, Australia.
E-mail address: kevin@maths.uwa.edu.au (K. Judd).

0167-2789/01/$ – see front matter © 2001 Published by Elsevier Science B.V.
PII: S0167-2789(01)00225-1

In practice, the ability to make good forecasts is limited both by uncertainty in the state and by imperfections in the model. In this paper, we consider the ideal situation where one has a perfect model. In this *perfect model scenario* one can make perfect forecasts for all time, *if* the true state of the system is known. Indeed, in this scenario a theorem of Takens [29] implies that if one could make perfect (noise free) observations of just a single scalar measurement of the system, then, for smooth finite-dimensional systems, one can (generically) obtain the true state from sufficiently many (but finite in number) observations. We consider here the case of noisy observation. In the second paper, we will extend the theory presented here to an imperfect model scenario [15].

In the first part of this paper we deal with the theory of indistinguishable states and in the second part we consider a new approach to data assimilation and state estimation. Atmospheric and oceanographic data assimilation exploit space–time variational methods for estimating a system's state [4,5,32]. In broad outline these methods take a time series of observations s_t, $t = 0, \ldots, p$, of a system with dynamics modeled by a function f, and attempt find an initial state $\hat{x}$ that minimizes the sum of squares error, [1]

$$E(x) = \sum_{t=0}^{p} (s_t - G(f^t(x)))^2, \tag{1}$$

where $s_t \in \mathrm{R}^k$, $x \in \mathrm{R}^d$, and G is an observer function that projects from state space R^d to observation space R^k.

One of the difficulties with this approach is that as the assimilation period p increases the repeated iteration of the nonlinear function f produces a complicated function $f^t(x)$ with the consequence that $E(x)$ has many local minima [1,23]. An alternative approach to forecasting is to look for *analogues* [19], i.e., to look through past records of the system's behavior (either observations or simulations) to find situations where the observations were like the present observations. The difficulty with this approach is that the recurrence time of the system may be so long [31] that no amount of observation, [2] or feasible amount of simulation, [3] would produce useful analogues of the present state and thus the construction of *perfect ensembles* [27,28] is untenable. In this paper, we present a new method of estimating the maximum likelihood state from observations that can produce model analogues of the current state without having to simulate the model for a long period. The method we suggest can be modified to allow for when the model f is imperfect or for dynamic noise (i.e., stochastic systems); these allowances are less easily accomplished in a variational approach.

The first part of this paper dealing with indistinguishable states has connections with Bayesian likelihood functions of state [1,3], but to our knowledge previous work has not revealed the intimate connection between dynamically invariant geometric structures and the support of the Bayesian density.

The second part of this paper has strong connections and similarities with nonlinear noise reduction [6,7,9,18]. The method we present could be viewed as (or even used for) nonlinear noise reduction, but our method differs from nonlinear noise reduction in both technical detail and underlying motivation. The results of the next section, however, are relevant to nonlinear noise reduction, because they imply that the goal of nonlinear noise reduction should be to obtain a distribution of states, not a single state.

2. Indistinguishable states

If one has a perfect model of a deterministic and finite-dimensional dynamical system, then the best forecast of the future is attained by determining the state of the system and evolving it forward. If knowledge of the state is

[1] In operational data assimilation there is usually an additional term $(x_0 - x_b)^2$, where x_b is some prior estimate of the initial state.
[2] The return time may exceed the life time of the system.
[3] The computation time may exceed the life time of the universe.

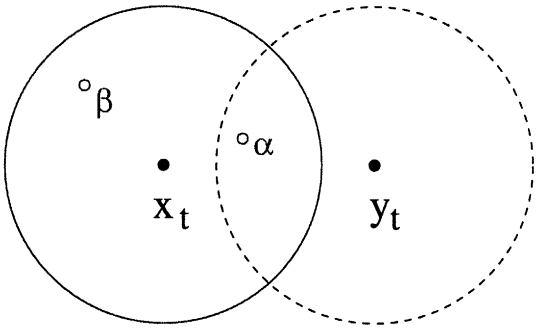

Fig. 1. When are two states x_t and y_t indistinguishable given an measurement error distributed by ρ? Suppose x_t is the true state of the system and y_t some other state. When the measurement error is bounded, then there is a bounded region about a state in which every possible observation of the state will fall. For the states x_t and y_t in the figure these regions are represented by circles centered on x_t and y_t. When an observation falls in the overlap of these regions (e.g., at α), then the states x_t and y_t are indistinguishable given this single observation. If the observation falls in the region about x_t, but outside the overlap with the region surrounding y_t (e.g., at β), then on the basis of this observation one can reject y_t being the true state, i.e., x_t and y_t are distinguishable given the observation. When the measurement error is unbounded, then states are never distinguishable on the basis of a single observation, although the probability of the states x_t and y_t are indistinguishable typically depends on the relative distances of the states from the observations.

clouded by measurement error, then one must estimate the state from what has been observed up to that moment. Is an exact estimation of the state is possible, even when given observations back to the beginning of time? The answer is no, as we now show.

For a deterministic system with state space $K \subseteq \mathrm{R}^d$ and initial state $x \in \mathrm{K}$ at $t = 0$, the state at a time t is $\Phi_t(x)$, where Φ is the evolution operator [8]. For convenience write $x_t = \Phi_t(x)$. In particular, $x_0 = \Phi_0(x) = x$. For notational convenience we will often drop the subscript and write x for x_0.

An observation s_t of the system state x_t at time t is corrupted by measurement error. Assume that $s_t = x_t + \epsilon_t$, where $\epsilon_t \in \mathrm{R}^d$ and has density ρ with respect to Lesbegue measure. [4] Assume also that observations are recorded at $t = 0, -1, -2, \ldots$, and that the ϵ_t are independent and identically distributed. Their mean need not be zero. The following results will generalize considerably from these assumptions, e.g., ρ can be time varying or state dependent and the measurement errors correlated, but such generalizations are avoided for clarity.

On the basis of a single observation s_t of x_t there can exist many states y_t each of which is indistinguishable from x_t, because of observational uncertainty; see Fig. 1. The joint probability density of x_t and y_t being indistinguishable is given by

$$\int \rho(s_t - x_t)\rho(s_t - y_t)\,\mathrm{d}s_t. \tag{2}$$

Define, by translating coordinates in the above,

$$g(b) = \int \rho(z)\rho(z - b)\,\mathrm{d}z, \tag{3}$$

where b corresponds to the separation between states $(y_t - x_t)$ and z is the actual measurement error. Normalize $g(b)$ to give

$$q(b) = \frac{g(b)}{g(0)}. \tag{4}$$

[4] Measurement error is usually taken to be Gaussian or bounded uniform, but neither singular nor fractal.

Observe that the joint probability (2) is then $g(y_t - x_t)$ and that the conditional probability that y_t is indistinguishable from the true state x_t is $q(y_t - x_t)$. The normalization implies that $q(0) = 1$ and thus, with probability one, x_t is indistinguishable from itself.

Given a time series of observations $s_t, t = 0, -1, -2, \ldots$, it follows (from the independence of the measurement error) that the probability that two trajectories x_t and $y_t, t = 0, -1, -2, \ldots$, are indistinguishable is given by

$$Q(y|x) = \prod_{t \leq 0} q(y_t - x_t). \tag{5}$$

From which immediately follows the theorem.

Theorem 1. *Given any time series of observations extending into the infinite past of the trajectory that terminates at x: if $Q(y|x) = 0$, then the states x and y are distinguishable with probability one.*

Some special cases of this theorem are the following.

- When the measurement error is bounded and for some $t = \tau$, x_t and y_t are separated by more than twice the bound, then $q(y_t - x_t) = 0$. (The circular regions shown in Fig. 1 would not overlap in this case.) It follows that $Q(y|x) = 0$, because at least one factor in the product (5) is zero. Thus such x and y will be distinguishable regardless of the realizations of the measurement error, since there is at least one moment in time (i.e., $t = \tau$) when a single observation will distinguish x from y.
- If the trajectories of x and y have a fixed separation and the measurement error is bounded, then $q(y_t - x_t)$ has a fixed value for all t. (The circular regions shown in Fig. 1 would have the same size overlap for all t.) Thus $Q(y|x) = 0$, as the infinite product (5) converges to zero, because it is a product of numbers strictly less than one. The interpretation in this case is that sooner or later there will be an observation that does not fall in the overlap region of Fig. 1, and such an observation will distinguish x from y. Note that it could happen that every measurement error places the observation in the overlap region, but there is zero probability of such a succession of measurement errors occurring.
- If the trajectories of x and y have a fixed separation, as in the above case, but the measurement errors are unbounded, such as Gaussian errors, then once again $q(y_t - x_t)$ has a fixed value for all t. Thus $Q(y|x) = 0$, since the infinite product (5) converges to zero. The interpretation in this case is that over the entire history of observations the evidence accumulates to distinguish the two trajectories that lead to x and y; even though no single observation distinguishes them, as happened in the previous cases.

Theorem 1 tells us when states can be distinguished under ideal circumstances. The important question now is thus: are there states that cannot be distinguished from the true state under ideal circumstances? If $Q(y|x) > 0$, then $Q(y|x)$ is the probability that the trajectories of x and y will not be distinguished, given a series of observations into the infinite past; here the probability is over all realizations of the measurement errors. Define

$$H(x) = \{y \in \mathrm{K} : Q(y|x) > 0\}, \tag{6}$$

i.e., $H(x)$ is the set of all possible states y that are indistinguishable from x given the entire history of the observations.

A more convenient alternative definition of $H(x)$ can be derived as follows. Define

$$h(b) = -\log q(b), \tag{7}$$

i.e., $h(y_t - x_t)$ is the likelihood that x_t and y_t cannot be distinguished given an observation, or alternatively, $h(y_t - x_t)$

is the information gained when an observation is made at time t. Now define

$$H(x) = \left\{ y \in \mathrm{K} : \sum_{t \le 0} h(y_t - x_t) < \infty \right\}. \tag{8}$$

This second equivalent form of the definition of $H(x)$ arises from considering sums of $h(y_t - x_t)$, rather than the products of $q(y_t - x_t)$ in Eq. (5).

Clearly, $x \in H(x)$, but if $H(x)$ is non-trivial (contains states other than x), then the "true" state x cannot be distinguished from the other states in $H(x)$. In the next section we demonstrate for three typical measurement error densities ρ, that $H(x)$ is non-trivial for typical nonlinear systems. It will be seen that the unstable set of x,

$$U(x) = \left\{ y \in \mathrm{K} : \lim_{\tau \to -\infty} \sup_{t \le \tau} \|y_t - x_t\| = 0 \right\}, \tag{9}$$

will play an important role, in that $H(x)$ is a subset of $U(x)$. It will also be seen that $H(x)$ can have a complicated nonlinear structure.

2.1. Examples of indistinguishability

To illustrate that $H(x)$ is typically non-trivial we consider when the density of measurement errors ρ is (i) Gaussian, (ii) bounded uniform, and (iii) a bounded non-uniform density.

Gaussian error density. Consider first when $d = 1$. Let

$$\rho(z) = \frac{1}{\sqrt{2\pi}\sigma} \, \mathrm{e}^{-z^2/2\sigma^2},$$

and so by (3), (4) and (7),

$$h(b) = \frac{b^2}{4\sigma^2}.$$

With (8) this means that $H(x)$ consists of all y such that $\sum_{t \le 0} (y_t - x_t)^2 < \infty$, which implies the trajectories x_t and y_t must converge in the past. Hence, $H(x)$ is a subset of the unstable set $U(x)$ of x; another way of putting this is that x and y have the same α-limit set [8]. Recall that for chaotic one-dimensional maps the divergence of trajectories is exponential on average, and so $H(x)$ almost certainly will be non-trivial.

When $d > 1$ define

$$\rho(z) = \left(\frac{1}{2\pi} \right)^{d/2} (\det A)^{1/2} \, \mathrm{e}^{-z^{\mathrm{T}} A z/2},$$

where A^{-1} is the covariance matrix of the measurement errors. One then has that

$$h(b) = \tfrac{1}{4} b^{\mathrm{T}} A b.$$

If the measurement errors are isotropic, then $A^{-1} = \sigma^2 I$, and $H(x)$ consists of all y such that $\sum_{t \le 0} \|y_t - x_t\|^2 < \infty$. When measurement errors are not isotropic, A will still be non-singular, and so it can be shown (by equivalence of norms) that $H(x)$ consists of all y such that $\sum_{t \le 0} \|y_t - x_t\|^2 < \infty$ in this case as well.

In general, for Gaussian measurement errors, $H(x)$ consists of all y such that $\sum_{t \le 0} \|y_t - x_t\|^2 < \infty$, which is a subset of the unstable set $U(x)$ of x. For differentiable systems, the divergence of the unstable set is typically exponential, so $H(x)$ is typically a sub-manifold of the state space K.

Uniform error. When $d = 1$,

$$\rho(z) = \begin{cases} \dfrac{1}{2r}, & |z| \le r, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$h(b) = -\log\left(1 - \frac{|b|}{r}\right) = \frac{|b|}{r} + \mathrm{O}\left(\frac{|b|}{r}\right)^2.$$

When $d > 1$ there are similar results for uniform density in a sphere or hypercube, and it can be shown that in general $H(x)$ consists of y such that $\sup_{t\le 0}\|y_t - x_t\| \le r$ and $\sum_{t\le 0}\|y_t - x_t\| < \infty$. This too is a subset of the unstable set of x, although more exclusive than the Gaussian case, because the density is bounded and because the asymptotics of the sum are more strict.

A non-uniform bounded error. Consider a generalized Epanechnikov density [11], defined for $d = 1$ by

$$\rho(z) = \begin{cases} \dfrac{15}{16r^5}(r^2 - z^2)^2, & |z| < r, \\ 0, & \text{otherwise} \end{cases}$$

for which

$$h(b) = -\log\left(1 - \frac{3b^2}{2r^2} + \frac{21b^4}{16r^4} - \frac{21b^5}{r^5} + \frac{105b^6}{2r^6} - \frac{393b^7}{8r^7} + \frac{315b^8}{16r^8} - \frac{23b^9}{8r^9}\right) = \frac{3b^2}{2r^2} + \mathrm{O}\left(\frac{b}{r}\right)^4.$$

There are similar results for $d > 1$ for the spherical or hypercube symmetries, and $H(x)$ is a subset of the unstable set of x again. Although asymptotically the sum is the same as the Gaussian density, $H(x)$ is more restricted because the density is bounded.

It is obvious from the preceding three examples that for an extensive variety of both bounded and unbounded densities, the set $H(x)$ will be non-trivial for differentiable and other systems. In particular, note that chaotic systems exhibit sensitivity to initial conditions, which is typically exponential divergence, and so one expects $H(x)$ to be non-trivial for almost all states of chaotic systems. It is not difficult to see that these results will generalize, e.g., not only to time varying and correlated measurement errors, but also to systems modeled by ordinary and partial differential equations.

2.2. Probability density of indistinguishable states

The function $Q(y|x)$ implies a probability measure on $H(x)$. Typically this probability measure has a maximum at x and decreases away from x along $H(x)$, but this is not always the case as we show in this section. For unbounded error densities, such as a Gaussian, $H(x)$ can be the entire unstable set of x. Note that typically the unstable set of x is spread about the entire attractor.

In this paper, we will use as an illustration the Ikeda map [13] in R^2 given by

$$f(u, v) = \begin{pmatrix} 1 + \mu(u\cos\theta - v\sin\theta) \\ \mu(u\sin\theta + v\cos\theta) \end{pmatrix}, \tag{10}$$

where $\theta = a - b/(1+u^2+v^2)$ with $a = 0.4$, $b = 6$ and $\mu = 0.83$. This map arises in the study of laser physics. We use it because it is chaotic and has a complex attractor that has dimension considerably more than one, and hence spread out in R^2. It is a more difficult map to deal with than, say, the Henon map.

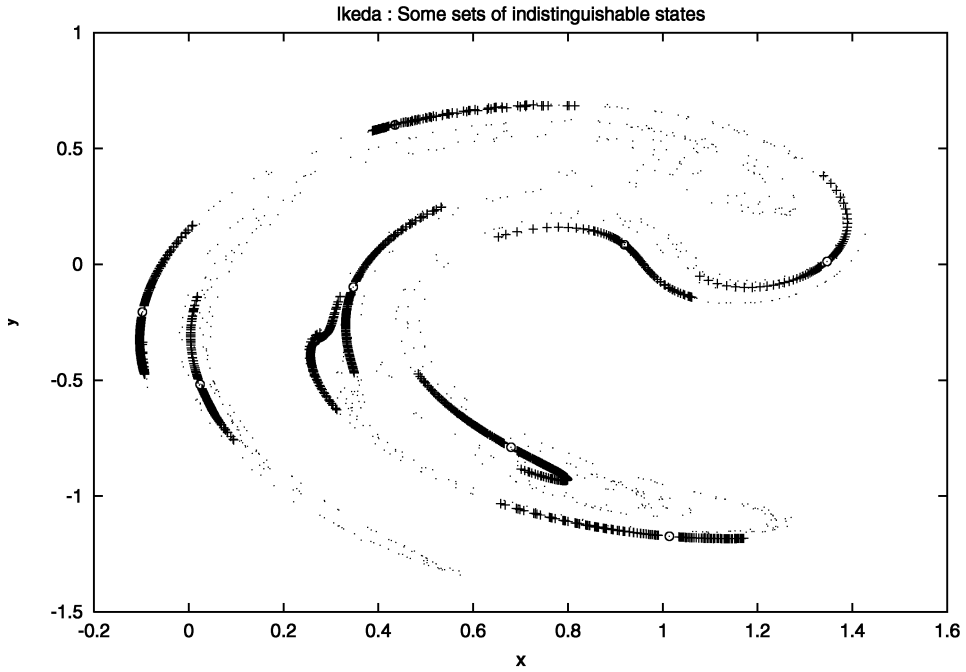

Fig. 2. The sets of indistinguishable states $H(x)$ calculated for nine states (marked with circles) of the Ikeda map [13] with measurement error that is Gaussian with a standard deviation $\sigma = 0.2$. The background of dots indicate the attractor of the system. The states marked with a circle are the selected true states x; those marked with plus signs are selected from $H(x)$ so as to indicate the extent of the set that contains at least 95% of the measure. See text for details on how the sets were calculated.

Fig. 2 shows finite samples of the sets $H(x)$ calculated for nine states x (marked with circles) of the Ikeda map (10), where the measurement error is Gaussian with a standard deviation $\sigma = 0.2$ — see next paragraph for a discussion of how the approximations were done. Observe that each of these approximations of $H(x)$ lie along the attractor, which is typically parallel to the unstable sets. Fig. 3 shows a close up of $H(0.264, -0.335)$ and the associated

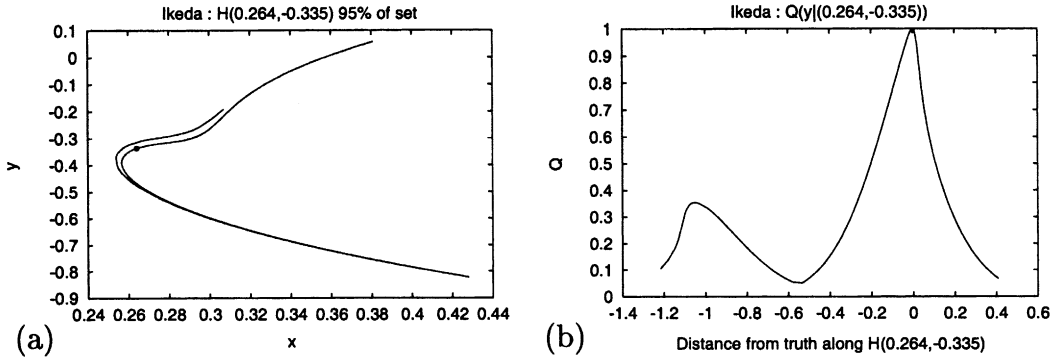

Fig. 3. Panel (a) is a close up of $H(0.264, -0.335)$ from Fig. 2 with the true state marked with a circle, and panel (b) the associated probability $Q(y|(0.264, -0.335))$ as a function of distance measured along the set $H(0.264, -0.335)$, where the positive direction is going up (increasing) from the true state $(0.264, -0.335)$. The next two peaks of the density have local maximums on the order of 10^{-6} and 10^{-8}. See text for details on how the set and the density were calculated.

probability $Q(y|(0.264, -0.335))$ as a function of distance measured along the set from $(0.264, -0.335)$. Observe that this set folds back upon itself and that the probability density is bimodal with asymmetric (skewed) peaks; in fact more than half of the sets $H(x)$ shown in Fig. 2 fold back like this and it appears to be typical behavior from this map. Note that the multi-modality occurs because there are trajectories shadowing the true trajectory which move away a short distance but are then brought back into the neighborhood of the true state; the return of trajectories like this is implied by the fold in the unstable set in which $H(x)$ lies. States that result from such trajectories have a high probability of being indistinguishable from the true state, because only a small number of fortuitously arranged measurement errors are needed. Note that for unbounded measurement error every close return to x of the unstable set $U(x)$ will produce a local maximum of the $Q(y|x)$ density, however, most of these close returns occur after the trajectory has made a series of deviations (on the length scale of the attractor) and will have a relatively small probability, i.e., the local maximum is small. For example, the next two peaks of the density beyond those shown in Fig. 3 have maximums on the order of 10^{-6} and 10^{-8}. Clearly, a significant fraction of $H(x)$ has been identified.

The sets $H(x)$ and the associated probabilities $Q(y|x)$ shown in Figs. 2 and 3 were calculated from a trajectory segment x_t, $t = -p, \ldots, 0$, where the final state $x_0 = x$ is the state for which $H(x)$ is desired. The initial point of this trajectory segment x_{-p} is perturbed and calling this perturbed point y_{-p}, the trajectory y_t, $t = -p, \ldots, 0$, is calculated and

$$Q_p(y|x) = \prod_{-p \le t \le 0} q(y_t - x_t). \tag{11}$$

Clearly, $\lim_{p\to\infty} Q_p(y|x) = Q(y|x)$, and for the Ikeda map calculations of Figs. 2 and 3 it was found that convergence was complete to several decimal places when $p = 32$. The extent of the set $H(x)$ was estimated by taking many perturbations of x_{-p} (in our calculations the perturbations formed an exponential spiral about x, but Guassian or bounded uniform perturbation of suitably chosen variance will work) until it was estimated that at least 95% of the Q measure of the set $H(x)$ had been indicated. If the size of the perturbation is too large most of the Q values are too small; if the size of the perturbations are too small $H(x)$ will not be covered. The size of the perturbation may, in some situations, be estimated using the largest Lyapunov exponent in the obvious manner.

The above method of sampling $H(x)$ will only give the approximate location of $H(x)$. Strictly speaking, $Q_p(y|x)$ will only converge for $y \in H(x)$, however, any state selected by the perturbation method we have described is almost certainly not in $H(x)$; in the Ikeda example $H(x)$ is a one-dimensional set but our candidate states are selected in two dimensions. If p is sufficiently large, then the distance of selections from $H(x)$ is insignificant compared to the separation between adjacent states of the finite sample.

2.3. Generalizations

The above analysis can be generalized in number of ways. Take the case of partial observation. Let $G : \mathrm{K} \to \mathrm{R}^k$ be a continuous observer function, so that, $s_t = G(x_t) + \epsilon_t$. Then the indistinguishable states $H(x)$ are those y such that $\sum_{t\le 0} h(G(y_t) - G(x_t)) < \infty$. Clearly, this set must include those states that are indistinguishable in the completely observable state case, but typically each of these points is broadened into a co-dimension k subset of K. A question of current interest in meteorology is to ask which k measurements should be made; the observations may also be adaptive, i.e., the selected components vary at each time step [10].

Many other generalizations within the prefect model scenario are straight forward, e.g., extension to ordinary differential equations, time- and state-dependent measurement error, etc. In a subsequent paper [15] we extend the theory of indistinguishable states to an imperfect model scenario.

3. Finding and forecasting with ensembles

Theorem 1 implies that the state of the system cannot be determined when $H(x)$ is non-trivial; internal consistency requires forecasts to take this uncertainty into account. Note that the maximum likelihood estimate of the state is of no special value, since its future behavior can be quite different from the true state. Consequently, one is forced either to find and evolve the density of indistinguishable states, or else approximate this distribution with a finite ensemble of states. We will consider only the construction of an ensemble. The method we suggest proceeds in two stages: first a maximum likelihood estimate of the true state is obtained, then, second, an ensemble is formed from a selection of states from the indistinguishable set of the maximum likelihood state. The maximum likelihood estimate of the true state could be obtained by variational methods as described in Section 1 (see Eq.(1)). As already mentioned variational approaches have short comings, such as multiple minima. We present now a new method of state estimation that avoids some of the drawbacks of variational approaches and is better suited to our goal of constructing an ensemble.

3.1. A new state estimation method

The principle idea of our new method is to take observations s_t of a trajectory and to relax these onto a near-by trajectory, without initially making any special effort to find the maximum likelihood trajectory.

Let $\mathrm{K} \subseteq \mathrm{R}^d$ (compact) be the state space and $f : \mathrm{K} \to \mathrm{K}$ (differentiable) be a perfect model of a discrete-time system. For any trajectory $x_t \in \mathrm{K}$, we have $x_{t+1} = f(x_t)$. If s_t are noisy observations of an unknown trajectory, then there exist δ_t such that

$$s_{t+1} - \delta_{t+1} - f(s_t - \delta_t) = 0. \tag{12}$$

For a time series of observations $s_t, t = 1, \ldots, p+1$ these are a set of equations in unknowns $\delta = (\delta_1, \ldots, \delta_{p+1}) \in \mathrm{R}^{(p+1)d}$, i.e., here are $(p+1) \times d$ parameters δ. There are only $p \times d$ equations, however, so the system of equations is under determined. [5] One could specify in advance δ_{p+1}, in which case the solution will be a trajectory that passes through the final state $s_{p+1} - \delta_{p+1}$, but this is not helpful given our interests.

For a finite segment of noisy observations, $s_t, t = 1, \ldots, p+1$, define

$$e_t = s_{t+1} - \delta_{t+1} - f(s_t - \delta_t), \tag{13}$$

and

$$L(\delta) = \frac{1}{2} \sum_{t=1}^{p} e_t^{\mathrm{T}} e_t. \tag{14}$$

One could attempt to find a solution to the system of Eq. (12) by solving the minimization problem

$$\min_{\delta} L(\delta). \tag{15}$$

Note that using the sum of squares assumes nothing about the distribution of the errors δ_t, it is simply a device to obtain a solution.

Theorem 2. *$L(\delta)$ has no local minima other than where $L(\delta) = 0$. In particular, if $Df(x)$ is of full rank for all $x \in \mathrm{K}$, i.e., if f is invertible, then $L(\delta)$ has no critical points other than on the subspace (see footnote 5) where $L(\delta) = 0$.*

[5] Indeed in δ-space $\mathrm{R}^{(p+1)d}$ there is a d-dimensional subspace with each point corresponding to a trajectory, it is parametrized by δ_1. If f is invertible, this is an embedding of K in δ-space.

Proof. Differentiating $L(\delta)$ we have,

$$\frac{\partial L}{\partial \delta_t} = \begin{cases} e_1^{\mathrm{T}} Df(s_1 - \delta_1), & t = 1, \\ -e_{t-1}^{\mathrm{T}} + e_t^{\mathrm{T}} Df(s_t - \delta_t), & 1 < t \leq p, \\ -e_p^{\mathrm{T}}, & t = p+1. \end{cases} \tag{16}$$

The critical points occur where $\partial L/\partial \delta_t = 0$ for each t. The $t = p+1$ equation is zero iff $e_p = 0$. Back-substitution of e_p into the $t = p$ equation shows that $e_p = 0$, and continued back-substitution shows $e_t = 0$ for all $0 \leq t \leq p$. Hence, $\partial L/\partial \delta_t = 0$ for each t iff $L(\delta) = 0$, and so $L(\delta)$ has no critical points other than where $L(\delta) = 0$. □

Theorem 2 implies that the minimization of $L(\delta)$ can be solved by local gradient descent, e.g.,

$$\dot{\delta} = -\frac{\partial L}{\partial \delta}\delta$$

for any initial condition when $Df(x)$ is of full rank, and almost all initial conditions otherwise. We have successfully used a stiff integrator [26], Powell's method [24], and stochastic descent [2]. We find that simplex descent [22] and (branch and bound) global optimizations [12,14], do not work well, however, and may wander to a solution $L(\delta) = 0$ that is not near the observed time series. Convergence results are discussed in detail in the next section.

One might argue that our new method for estimating the state (by finding a near-by trajectory by gradient descent of $L(\delta)$) has no great advantage over a variational approach. That is, one may argue that the problem of multiple minima of the variational approach [1,23] has been replaced by an infinite subspace of solutions given by $L(\delta) = 0$. Although this is true, our counter argument is that because the true state is unknowable (it is indistinguishable from many others) it is better to have a method that guarantees a solution that is suitable for constructing ensemble estimates of the state (as we will show), than use a variational approach that cannot guarantee such an outcome.

3.2. Properties of gradient descent solutions

In this section, the trajectories obtained by minimizing $L(\delta)$ by gradient descent are shown to have useful properties for ensemble forecasting. Define the stable set $S(x)$ of x in the analogous way to the unstable set $U(x)$ in Eq. (9):

$$S(x) = \left\{ y \in \mathrm{K} : \lim_{\tau \to \infty} \sup_{t \leq \tau} \|y_t - x_t\| = 0 \right\}. \tag{17}$$

We propose the following dictum. [6]

Dictum 1. *Let $S(x_1) \subseteq \mathrm{K}$ and $U(x_{p+1}) \subseteq \mathrm{K}$ be the stable and unstable sets of the end-points of a segment of trajectory. Then the trajectory obtained by the minimization of $L(\delta)$ by gradient descent given noisy observations along this segment will have end-points that are close to $S(x_1)$ and $U(x_{p+1})$.*

[6] A dictum is a formal statement of opinion; while it is true for typical behavior of typical systems, counter examples can be constructed. Typically, a dictum can be turned into a theorem by adding sufficiently many restrictions (in the present case: hyperbolicity, that the trajectory be a typical trajectory of the attractor (Gibbs state), and so on). Turning the dictum into a theorem at present, however, does not seem warranted since doing so provides little illumination of general behavior. What is of interest here is not the narrowly true case but the useful observation. On the other hand, we anticipate publishing a theorem [25] (see also [33]).

Justification. $L(\delta) = 0$ implies that $\hat{x}_t = s_t - \delta_t$ is a trajectory of the system defined by f. This trajectory need not be the same as the true trajectory x_t. Let $\hat{x}_t = x_t + \eta_t$, i.e., η_t is the deviation of $\hat{x}_t$ from the true trajectory. Substituting for $s_t - \delta_t = \hat{x} = x_t - \eta_t$ in Eq. (13) gives

$$e_t = x_{t+1} + \eta_{t+1} - f(x_t + \eta_t). \tag{18}$$

Assuming that the η_t are small relative to the nonlinearity of f implies that

$$e_t = x_{t+1} + \eta_{t+1} - f(x_t) - A_t\eta_t + \mathrm{O}(\eta)^2 \tag{19}$$

$$= \eta_{t+1} - A_t\eta_t + \mathrm{O}(\eta)^2, \tag{20}$$

where $A_t = Df(x_t)$. But $L(\delta) = 0$ implies $e_t = 0$ and so the above implies $\eta_{t+1} = A_t\eta_t + \mathrm{O}(\eta)^2$, from which we obtain

$$\eta_t = A_{t-1}A_{t-2}\cdots A_1\eta_1 + \mathrm{O}(\eta)^2. \tag{21}$$

If f is invertible, then the A_t are invertible too, and so we also obtain

$$\eta_t = A_t^{-1}A_{t+1}^{-1}\cdots A_p^{-1}\eta_{p+1} + \mathrm{O}(\eta)^2. \tag{22}$$

If f also happens to define a hyperbolic system, then there is a splitting of A_t into stable and unstable blocks. Eqs. (21) and (22) imply that (relative to other η_t) η_1 aligns with the stable set $S(x_1)$ and η_{p+1} aligns with the unstable set $U(x_{p+1})$.

Of course, the situation is somewhat more complex than this because the orientations of the stable and unstable eigenspaces generally do not always align with those of the following period except in hyperbolic systems. Furthermore, solving $L(\delta) = 0$ by gradient descent would also seem to require $Df(s_t - \delta_t) \approx Df(x_t)$ at each stage of the gradient descent. It will be seen later that in practice it seems the method generally works even when neither of these conditions are met.

The behavior described in Dictum 1 has been observed in other nonlinear noise reduction techniques [7,9].

Fig. 4 indicates maximum likelihood estimates of the end-points of trajectory segments resulting from gradient descent of $L(\delta)$ for nine noisy observations for various states of the Ikeda map. The algorithm for finding maximum likelihood estimates is a modification of the algorithm just discussed and is described in the next section. For the Ikeda map the differences between the maximum likelihood estimates, and the estimates obtained with the algorithm just discussed are almost indistinguishable on the scale of the figure.

These results displayed in Fig. 4 also support the dictum, because the states obtained appear to lie close to the stable and unstable sets. It is worth noticing that the estimates for the first point (plus signs) are not as close to the stable set as the estimates of the last point (crosses) are to the unstable set. This can be attributed entirely to not having followed the gradient descent long enough to attain the final $L(\delta) = 0$ condition; this point is taken up in the discussion of Fig. 5 to follow. That there is slower convergence in the stable direction can be understood as a consequence of the magnitude of the stable and unstable Lyapunov exponents (see justification of Dictum 1).

Note that in Fig. 4 the estimated final state (plus signs) reflect the distribution on $H(x)$ according to the measure induced by probability of indistinguishablity Q; when compared with Fig. 2 there is good agreement with the location of $H(x)$ for each state.

Fig. 5 shows the evolution of δ during a gradient descent minimization using a stiff integration routine for one of the worst cases in Fig. 4. In this case the stable component has not fully converged. At the 80th control point, nominal integration time $\approx 10^2$, all the δ are close to their final values, but when the descent was

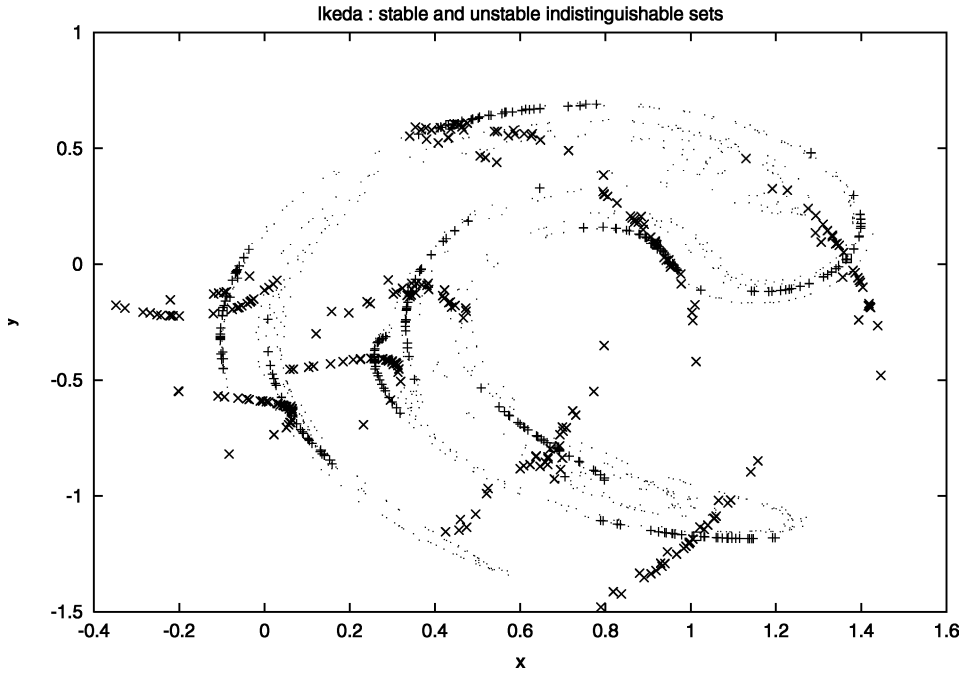

Fig. 4. Maximum likelihood state estimation is illustrated here by the Ikeda map [13]. The observation error was Gaussian of mean 0 and standard deviation 0.2, about a one tick interval. The background of dots indicate the attractor of the system, and the squares (sometimes obscured) the states chosen to illustrate method. For each chosen state we consider what happens when it is the first or the last state of the trajectory segment of nine points; 30 separate observations of the trajectory segment in each case. The crosses indicate the estimated state when the first state and the plus signs the last state of the trajectory segment. See text for a detailed discussion of these results, but note that the plus signs spread out along the unstable sets of the true states and should be distributed on the set of indistinguishable states according the measure induced by Q; compare this figure with Fig. 2 which uses the same selection of states.

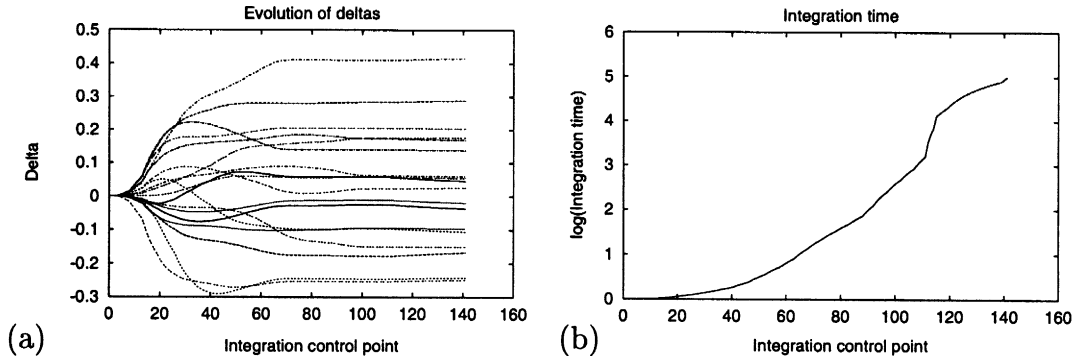

Fig. 5. Evolution of deltas in a gradient descent minimization for the uppermost point of Fig. 4. The nine point trajectory segment gives a total of 18 deltas. The gradient descent was performed with the `ode15s` stiff integration routine [26] of `Matlab`. In panel (a) the deltas are plotted as a function of the integration control points, which are not uniformly spaced in time. The equivalent nominal integration time of these control points are shown in panel (b); note this is base-10 logarithm time. In this example the stable component is only around 90–95% converged at the termination of the descent; the example chosen is a worst case, typically convergence is faster than illustrated.

terminated after a nominal time of 10^5, some δ components are clearly still changing. A more detailed analysis of choices of trajectory segment length, descent time and accuracy remains to be done. In comparison Powell's minimization algorithm [24], or modifications of it, may be faster. We have obtained more accurate and consistent results with a stiff integrator, but this may be because the stopping criteria for Powell's method have not been optimized. Similarly, stochastic minimization gives good low accuracy results over similar computation times; obtaining higher accuracy results would most likely require introducing an cooling schedule [20].

3.3. Maximum likelihood trajectories

The trajectories obtained by gradient descent minimization are not maximum likelihood estimates of the true trajectory, but this short coming is easily corrected as we will now show. Finding the maximum likelihood trajectories when the measurement error is Gaussian is equivalent to finding δ that minimize $\delta^{\mathrm{T}}\delta$ subject to the constraint $L(\delta) = 0$, or as it is more usually implemented, subject to the constraints $e_t = 0$ for $t = 1, \ldots, p$. One could attempt to solve this nonlinear optimization problem (with nonlinear constraints) using standard Lagrangian, penalty function, boundary function, or interior point methods. Good general purpose software for this purpose is not widely available, and the few that we have access to, and have tried, have not performed as well as the simple, easily implemented method we describe below. (The general purpose algorithms often fail to converge or do so very slowly.) However, we have evidence that a purpose designed nonlinear optimizer that exploits structures of the problem is as accurate as gradient descent but faster and more efficient in terms of function evaluations, and consequently might be the best option for large scale problems [17].

Finding the maximum likelihood trajectories when the measurement error is Gaussian is equivalent to finding δ such that $L(\delta) = 0$ and $\delta^{\mathrm{T}}\delta$ is as small as possible. One might be tempted to solve

$$\min_{\delta} L(\delta) + a\delta^{\mathrm{T}}\delta, \tag{23}$$

where $a > 0$, i.e., to attempt to simultaneously minimize $L(\delta)$, as in (15), and $\delta^{\mathrm{T}}\delta$. This is not advised, because the minimum may occur where $L(\delta) \neq 0$, i.e., the solution is not a trajectory. One can, however, obtain a maximum likelihood trajectory to any required accuracy, either by alternately solving by gradient descent $\min_\delta L(\delta)$ and $\min_\delta L(\delta) + a\delta^{\mathrm{T}}\delta$, with a small, using the final δ of one minimization as the initial condition of the next, or by sequentially solving $\min_\delta L(\delta) + a_n\delta^{\mathrm{T}}\delta$ with $a_n \to 0$. For a general error density $\rho(\delta)$, one should solve

$$\min_{\delta} L(\delta) - a\sum \log(\rho(\delta)).$$

The maximum likelihood states shown in Fig. 4 were estimated by minimization of (23) with $a = 10^{-4}$ followed by minimization of $L(\delta)$.

Fig. 6 shows relative RMS differences between maximum likelihood estimates and a true trajectory; the RMS error is relative to the measurement error, which is Gaussian with standard deviation $\sigma = 0.2$. Contrasting the results for various segment lengths p reveals that the RMS errors at time step 0 appear to have largely converged by $p \geq 6$. The asymptotic RMS error is some 50% of the noise σ, bearing in mind this variance is theoretically restricted to the one-dimensional unstable set, which is seen to have been roughly attained from Fig. 4. This outcome is better than what has previously been proposed as the best possible noise reduction [23]. The addition reduction may result because the previous cited calculation were based on local linearization estimates, whereas our results are based a fully nonlinear theory, i.e., nonlinearity of the system can provide addition information about trajectories that is exploited in the noise reduction.

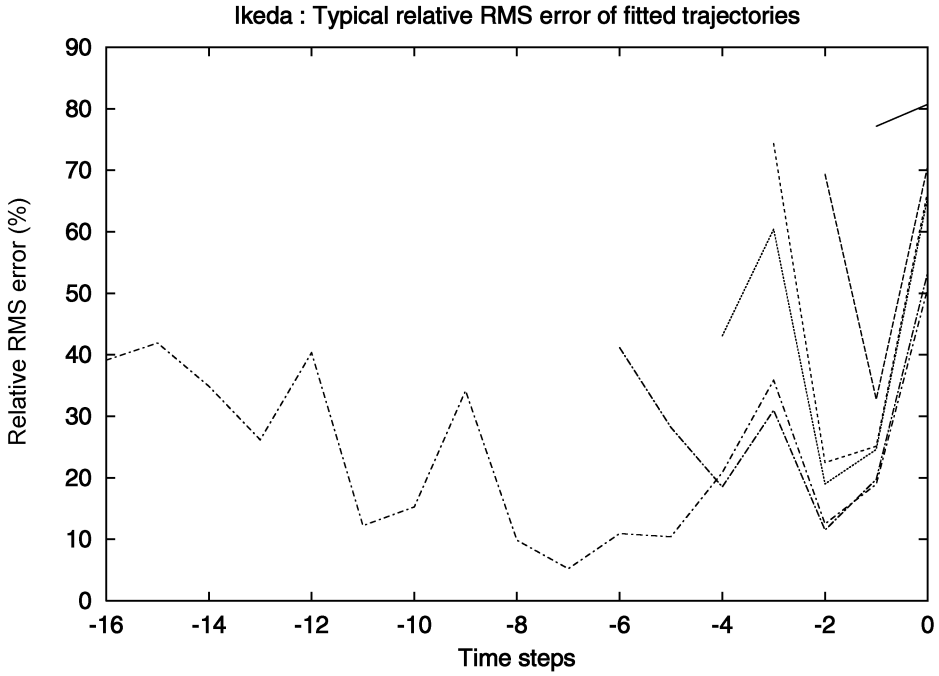

Fig. 6. The relative root mean square differences between maximum likelihood estimates of observed trajectories and a true trajectory; the RMS error is relative to the measurement error, which is Gaussian with standard deviation $\sigma = 0.2$. The true trajectory used here terminates near (0.5, 0.5) and is shown in Fig. 4. This graph plots the RMS difference for each point along the trajectory segment ($t = -p$ being the start and $t = 0$ the final point) and was obtained for segments with $p = 1, 2, 3, 4, 6, 16$ with a sample of 60 observations of each segment.

3.4. Ensembles of indistinguishable states

We are now in a position to suggest how to draw a sample from the distribution of indistinguishable states to form an ensemble. This is relevant to forecasting: given an ensemble of indistinguishable states and their relative likelihood Q, we can maintain the uncertainty inherent in our observation in our forecast.

Recall that our proposed method of obtaining an ensemble is to first obtain a maximum likelihood estimate $\hat{x}$ of the true state x and then select ensemble members from the indistinguishable set of the maximum likelihood estimate $H(\hat{x})$. We desire that the ensemble constructed in this way "capture" truth, i.e., we require that there is a good chance that $x \in H(\hat{x})$. Now, there is an essential symmetry that $x \in H(\hat{x})$ iff $\hat{x} \in H(x)$, and so the following dictum provides a justification that $x \in H(\hat{x})$.

Dictum 2. *Let s_t, $t = 0, -1, -2, \ldots$, be an observation of the trajectory of x with errors $\epsilon_t = s_t - x_t$ having density ρ with mean zero and $\rho(\epsilon)$ a decreasing function of $\|\epsilon\|$. Let $\Psi_p(s) = \Psi_p(s_{-p}, \ldots, s_0)$ map the last p observations to the maximum likelihood final state given these observations. Then*

$$\lim_{p\to\infty} \Psi_p(s) \in H(x)$$

with probability one.

Justification. Referring back to the justification of Dictum 1, we have that $x_t + \eta_t = s_t - \delta_t$. Using Eqs. (21) and (22), with slightly different indexing, gives

$$\delta_t = s_t - x_t + A_{t-1}A_{t-2}\cdots A_{-p}\eta_{-p} + \mathrm{O}(\eta)^2, \tag{24}$$

or

$$\delta_t = s_t - x_t + A_t^{-1}A_{t+1}^{-1}\cdots A_1^{-1}\eta_0 + \mathrm{O}(\eta)^2. \tag{25}$$

These imply that each δ_t is composed of two parts: the actual observation error $\epsilon_t = s_t - x_t$, and the deviation from truth (indistiguishablity component) parameterized by η_{-p} or η_0. These two components are not independent, but to a first approximation the minimization of, e.g., $\delta^{\mathrm{T}}\delta$ involves a fixed residual $\epsilon^{\mathrm{T}}\epsilon$ and a term involving η_{-p} or η_0, ignoring the mixed-product terms. Clearly, the term involving η_{-p} or η_0 attains its minimum where η_{-p} is aligned with $S(x_{-p})$ and η_0 is aligned with $U(x_0)$ for large p. This assumes hyperbolicity. Of course, the δ_t are determined by the realization of the noise ϵ_t in a complex way, however, if the noise has zero mean, then δ_t has expected value equal to η_t as given by Eqs. (21) and (22).

We propose two methods of selecting states from the indistinguishable set $H(x)$ to form an ensemble: (i) renoising the maximum likelihood trajectory estimate, and (ii) perturbation of an initial state, similar to that used for Fig. 3.

Given that the error density ρ is known, one can obtain an ensemble approximation of the density of indistinguishable states in $H(x)$ by renoising a maximum likelihood estimate of the trajectory. Given a time series of observations s, obtain the maximum likelihood estimate $y = \Psi_p(s)$, where p should be large enough that variations of s_{-p}, and hence y_{-p}, do not significantly effect y_0, i.e., y_{-p} and y_0 are largely independent as far as Ψ is concerned. Generate simulated observations S (surrogates) from y by adding ρ-random-variates to y, and for each of these obtain $Y = \Psi_p(S)$; the Y and initial y form the ensemble. The ensemble so produced is a selection from $H(y)$ by the probability measure induced by Q, in the limit of $p \to \infty$. (Note that one could alternatively renoise the observations s, instead of renoising y, which adds a certain amount of bias.)

The alternative method (which is likely to be more efficient) is to use a method similar to that used in Fig. 2 to calculate $H(x)$ and $Q(y|x)$. That is, one takes a segment of the maximum likelihood trajectory, y_t, $t = -p, \ldots, 0$ and generates a perturbation z_{-p} of the initial point y_{-p} in order to calculate the probability $Q_p(z|y)$. The states $z = z_0$ and probabilities $Q_p(z|y)$ form a weighted ensemble. There may be some benefit in perturbing a state that is not the initial state of the maximum likelihood trajectory, but rather a little way from the beginning of the trajectory segment where the uncertainty in both the stable and unstable direction is less, but we have not yet pursued this in detail.

Note that by making the ensemble large enough it will straddle the true state x. When Dictum 2 holds, the maximum likelihood estimate of the state y lies in $H(x)$ and by symmetry $x \in H(y)$. As the ensemble size is increased there will be states in the ensemble that approximate x arbitrarily closely. For a more detailed discussion of how large an ensemble should be see [16].

4. Conclusion

In this paper, we have demonstrated for typical dynamical systems, in particular, any system displaying sensitivity to initial conditions, that it is not possible to determine the state of the system when there is measurement error, even when given arbitrary amounts of historical observation and a perfect model. This is contrary to what one might have guessed, i.e., this is unlike many statistics (like means and variances) that converge to the true value as more and more data is collected. Instead we have shown there is a set of states that are indistinguishable from the true state given the observations; there is a probability distribution on the indistinguishable states.

Indistinguishable states imply that a probabilistic approach to forecasting is required even with perfect models of deterministic systems. We have considered an ensemble forecasting approach, and described a new method of ensemble construction that first obtains the maximum likelihood estimate of the true state, and then determines states which are indistinguishable from this maximum likelihood state along with their relative probability. Our new approach to estimating the maximum likelihood state has advantages over variational approaches. Ensemble forecasting has been operational at both American and European weather centers for sometime [21,30]; it follows from our results that this approach is required in nonlinear noise reduction, in prediction, and in forecast evaluation of lower-dimensional dynamical systems as well.

Acknowledgements

This research has been supported by an ONR grant (ONR Predictability DRI N00014-99-1-0056) and an Australian Research Council grant. KJ thanks the Oxford Centre for Industrial and Applied Mathematics for its hospitality. This paper has benefited from comments and suggestions of H. Abarbanel, D. Broomhead, J. Hansen, J. Huke, P. McSharry, A. Mees, M. Muldoon, D. Ridout, and K. Skouras. D. Ridout provided significant suggests to simplify the justifications of Dictums 1 and 2 and anticipates providing theorems for Dictums 1 and 2 [25] (see also [33]).

References

[1] L.M. Berliner, Likelihood and Bayesian prediction of chaotic systems, J. Am. Statist. Assoc. 86 (416) (1991) 938–952.
[2] C.G.E. Boender, H.E. Romeijn, Stochastic methods, in: R. Horst, P.M. Pardalos (Eds.), Handbook of Global Optimization, Kluwer Academic Publishers, Dordrecht, 1995, pp. 829–869.
[3] M. Casdagli, S. Eubank, J.D. Farmer, J. Gibson, State space reconstruction in the presence of noise, Physica D 51 (1991) 52.
[4] P. Coutier, Variational methods, J. Meteorol. Soc. Jpn. B 75 (1) (1997) 211–218.
[5] R. Daley, Atmospheric Data Analysis, Cambridge University Press, Cambridge, 1991.
[6] M. Davies, Noise reduction by gradient descent, Int. J. Bifur. Chaos 3 (1) (1992) 113–118.
[7] M. Davies, Noise reduction schemes for chaotic time series, Physica D 79 (1994) 174–192.
[8] J. Guckenheimer, P. Holmes, Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields, Vol. 42, Springer, New York, 1983.
[9] S.M. Hammel, A noise-reduction method for chaotic systems, Phys. Lett. A 148 (1990) 421–428.
[10] J.A. Hansen, L.A. Smith, The role of operational constraints in selecting supplementary observations, J. Atmos. Sci. 57 (2000) 2859–2871.
[11] W. Härdle, Applied Nonparametric Regression, Economic Society Monographs, Vol. 19, Cambridge University Press, Cambridge, 1990.
[12] W. Huyer, A. Neumaier, Global optimization by multilevel coordinate search, Technical Report, Institut für Mathematik, Universität Wien, 1999.
[13] K. Ikeda, Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system, Opt. Commun. 30 (1979) 257.
[14] D.R. Jones, C.D. Perttunen, B.E. Stuckman, Lipschitzian optimization without the Lipschitz constant, J. Opt. Theor. Appl. 79 (1993) 157–181.
[15] K. Judd, L.A. Smith, Indistinguishable states II: an imperfect model scenario, in preparation.
[16] K. Judd, L.A. Smith, On bounding boxes and ensembles, in preparation.
[17] D. Kilminster, K. Judd, Fitting models to systems with observational and dynamical noise, in preparation.
[18] S.P. Lalley, Removing the noise from chaos plus noise, in: A.I. Mees (Ed.), Nonlinear Dynamics and Statistics, Proceedings of the Isaac Newton Institute Workshop, Birkhauser, Boston, 2000.
[19] E.N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues, J. Atmos. Sci. 26 (1969) 636–646.
[20] M. Lundy, A.I. Mees, Convergence of an annealing algorithm, Math. Prog. 34 (1986) 111–124.
[21] F. Molteni, R. Buizza, T.N. Palmer, T. Petroliagis, The ECMWF ensemble prediction system: methodology and validation, Quart. J. R. Meteorol. Soc. 122 (1996) 73–120.
[22] J.A. Nelder, R. Mead, Comput. J. 7 (1965) 308.
[23] C. Pires, R. Vautard, O. Talagrand, On extending the limits of variational assimilation in nonlinear chaotic systems, Tellus A 48 (1996) 96–121.

[24] M.J.D. Powell, TOLMIN: a Fortran package for linearly constrained C optimization calculations, Technical Report NA2, DAMTP, University of Cambridge, Cambridge, 1998.
[25] D. Ridout, K. Judd, Convergence properties of noise reduction by gradient descent, in preparation.
[26] L.F. Shampine, M.W. Reichlet, The Matlab ODE suite, Technical Report, The MathWorks, Inc., Prime Park Way, Natick, MA, 1995.
[27] L.A. Smith, Accountability in ensemble prediction, in: Proceedings of the ECMWF Workshop on Predictability, Vol. 1, ECMWF, Shinfield Park, Reading, UK, 1996, pp. 351–368.
[28] L.A. Smith, C. Zimmerman, K. Fraedrich, Uncertainty dynamics and predictability in chaotic systems, Quart. J. R. Meteorol. Soc. 125 (1999) 2855–2886.
[29] F. Takens, Detecting strange attractors in turbulence, in: D.A. Rand, L.S. Young (Eds.), Dynamical Systems and Turbulence, Vol. 898, Springer, Berlin, 1981, pp. 365–381.
[30] Z. Toth, E. Kalnay, Ensemble forecasting at NMC: the generation of perturbations, Bull. Am. Meteorol. Soc. 74 (12) (1993) 2317–2330.
[31] H.M. van der Dool, Searching for analogues, how long must we wait? Tellus A 46 (1994) 314–324.
[32] C. Wunsch, The Ocean Circulation Inverse Problem, Cambridge University Press, Cambridge, 1996.
[33] D. Ridout, Convergence properties of noise reduction by gradient descent, MSc Thesis, Department of Mathematics and Statistics, University of Western Australia, 2001.